A Majorize-Minimize line search algorithm for barrier functions

Emilie Chouzenoux, Saïd Moussaoui, Jérôme Idier

Rapport technique IRCCyN

A Majorize-Minimize line search algorithm for barrier functions

Emilie Chouzenoux · **Saïd Moussaoui** ·
Jérôme Idier

September 7, 2010

Abstract Criteria containing a *barrier* function i.e., an unbounded function at the boundary of the feasible solution domain are frequently encountered in the optimization framework. When an iterative descent method is used, a search along the line supported by the descent direction through the minimization of the underlying scalar function has to be performed at each iteration. Usual line search strategies use an iterative procedure to propose a stepsize value ensuring the fulfillment of sufficient convergence conditions. The iterative scheme is classically based on backtracking, dichotomy, polynomial interpolations or quadratic majorization of the scalar function. However, since the barrier function introduces a singularity in the criterion, classical line search procedures tend to be inefficient. In this paper we propose a majorization-based line search strategy by deriving a nonquadratic form of a majorant function well suited to approximate a criterion containing a barrier term. Furthermore, we establish the convergence of classical descent algorithms when this strategy is employed. The efficiency of the proposed line search strategy is illustrated by means of numerical examples in the field of signal and image processing.

Keywords Descent optimization methods · barrier function · line search · majorize-minimize algorithm · convergence

1 Introduction

The aim of this paper is to address optimization problems that read

$$\min_{\boldsymbol{x}} \{F(\boldsymbol{x}) = P(\boldsymbol{x}) + \mu B(\boldsymbol{x})\}, \quad \mu > 0 \tag{1}$$

This work was supported by the OPTIMED project of the French National Research Agency (ARA MDMSA).
The authors thank Gilles Aubert and Jean-Pierre Dussault for critical reading of an early draft version.

IRCCYN, CNRS UMR 6597
1, rue de la Noë, BP 92101, F-44321 Nantes Cedex 03, France
E-mail: emilie.chouzenoux, said.moussaoui, jerome.idier@irccyn.ec-nantes.fr

where $\boldsymbol{x} \in \mathbb{R}^n$, B is a *barrier* function having its gradient unbounded at the boundary of the strictly feasible domain

$$\mathscr{C} = \{\boldsymbol{x} | C_i(\boldsymbol{x}) > 0,\, i = 1, ..., m\}$$

and P is differentiable on $\mathscr{C}$. We consider the case of linear constraints $C_i(\boldsymbol{x}) = \boldsymbol{a}_i^T \boldsymbol{x} + \rho_i$ with $\boldsymbol{a}_i \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$, $\rho_i \in \mathbb{R}$ and barrier functions that read

$$B(\boldsymbol{x}) = \sum_{i=1}^{m} \psi_i(C_i(\boldsymbol{x})) \tag{2}$$

with ψ_i taking one of the following forms:

$$\psi_i(u) = -\kappa_i \log u,\, \kappa_i > 0 \tag{3}$$

$$\psi_i(u) = \kappa_i u \log u,\, \kappa_i > 0 \tag{4}$$

$$\psi_i(u) = -\kappa_i u^r, \quad r \in (0,1),\, \kappa_i > 0 \tag{5}$$

so that the minimizers $\boldsymbol{x}^*$ of F fulfill $C_i(\boldsymbol{x}^*) > 0$.

A large family of optimization methods to solve (1) are based on iteratively decreasing the criterion by moving the current solution $\boldsymbol{x}_k$ along a direction $\boldsymbol{d}_k$,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \tag{6}$$

where $\alpha_k > 0$ is the *stepsize* and $\boldsymbol{d}_k$ is a *descent direction* i.e., a vector satisfying $\nabla F(\boldsymbol{x}_k)^T \boldsymbol{d}_k < 0$. Such iterative descent methods consist in alternating the construction of $\boldsymbol{d}_k$ and the determination of α_k (*line search*). While the direction is computed using the criterion properties (gradient, Hessian) at the current value $\boldsymbol{x}_k$, the line search is performed by minimizing the scalar function $f(\alpha) = F(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$. Some iterative methods do not require the line search step since the direction is calculated such that the optimal value of α_k would be equal to one (e.g., trust region algorithms ([6]), subspace optimization ([36,28]) or variable metric algorithms ([10,14])). Our analysis does not cover this family of methods.

Usual line search strategies perform an inexact minimization of f and propose a stepsize value that ensures the convergence of the descent algorithm ([31]). Typically, an iterative procedure generates a series of stepsize values until the fulfillment of sufficient convergence conditions such as Wolfe and Goldstein conditions ([26,31]). The iterative scheme is classically based on backtracking or dichotomy and more sophisticated procedures involve polynomial interpolations of the scalar function. Another alternative is to use quadratic majorizations of the scalar function leading to stepsize formulas guaranteeing the overall algorithm convergence ([37,22]). However, since the barrier function in problem (1) has a singularity at the boundary of $\mathscr{C}$, the derivative of the scalar function is unbounded which makes polynomial interpolation-based strategies inefficient ([27]) and quadratic majorization unsuited.

In this paper a majorization-based line search is firstly proposed by deriving a nonquadratic form of a majorant function well suited to approximate a criterion containing a barrier term. Secondly, convergence results are obtained for classical descent algorithms when this strategy is applied. The rest of this paper is organized as follows: After introducing the framework of the optimization problem in §2, we explain in §3 why special-purpose line search procedures are called for when dealing with barrier functions. A suitable line search strategy based on majorization is then proposed in §4. §5 gives the properties of the resulting stepsize series and §6 presents the convergence results when the proposed line search is associated with classical descent algorithms. §7 illustrates the efficiency of the proposed line search strategy through numerical examples in the field of signal and image processing.

2 Preliminaries

Assumption 1 *Let $\mathscr{V}$ be a neighborhood of the level set $\mathscr{L}_0 = \{\boldsymbol{x}|F(\boldsymbol{x}) \leqslant F(\boldsymbol{x}_0)\}$. $\mathscr{V}$ is assumed bounded. Moreover, $F : \mathbb{R}^n \to \mathbb{R}$ is differentiable on $\mathscr{V}$ and $\nabla F(\boldsymbol{x})$ is Lipschitz continuous on $\mathscr{V}$ with the Lipschitz constant $L > 0$:*

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathscr{V}$$

The first part of the assumption is not a restrictive condition since it holds if F is coercive, that is:

$$\lim_{\|\boldsymbol{x}\| \to +\infty} F(\boldsymbol{x}) = +\infty$$

According to Assumption 1, there exists $\eta > 0$ such that

$$\|\nabla F(\boldsymbol{x})\| \leqslant \eta, \forall \boldsymbol{x} \in \mathscr{V} \tag{7}$$

Moreover, because the gradient of B is unbounded at the boundary of $\mathscr{C}$, (7) leads to the existence of $\varepsilon_0 > 0$ such that

$$C_i(\boldsymbol{x}) \geqslant \varepsilon_0, \forall \boldsymbol{x} \in \mathscr{V}, \forall i = 1, \ldots, m, \tag{8}$$

and the boundedness assumption on $\mathscr{V}$ implies that there exists $M > 0$ such that

$$C_i(\boldsymbol{x}) \leqslant M, \forall \boldsymbol{x} \in \mathscr{V}, \forall i = 1, \ldots, m. \tag{9}$$

Assumption 2 *Assumption 1 holds and F is convex on $\mathscr{V}$: for every $(\boldsymbol{x}, \boldsymbol{y}) \in \mathscr{V}$ we have*

$$F(\omega \boldsymbol{x} + (1-\omega)\boldsymbol{y}) \leqslant \omega F(\boldsymbol{x}) + (1-\omega)F(\boldsymbol{y}), \forall \omega \in [0, 1]$$

Assumption 3 *Assumption 1 holds and F is strongly convex on $\mathscr{V}$: there exists $\lambda > 0$ such that*

$$\left[\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{x}')\right]^T (\boldsymbol{x} - \boldsymbol{x}') \geqslant \lambda \|\boldsymbol{x} - \boldsymbol{x}'\|^2, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathscr{V}$$

Definition 1 Let $\{\boldsymbol{M}_k, k = 1, \ldots, K\}$ a set of symmetric matrices. $\{\boldsymbol{M}_k\}$ has a *non-negative bounded spectrum* with bounds $(\nu_1^{\mathscr{M}}, \nu_2^{\mathscr{M}}) \in \mathbb{R}$ if for all k,

$$0 \leqslant \nu_1^{\mathscr{M}} \leqslant \frac{\boldsymbol{x}^T \boldsymbol{M}_k \boldsymbol{x}}{\|\boldsymbol{x}\|^2} \leqslant \nu_2^{\mathscr{M}}, \forall \boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\} \tag{10}$$

Moreover, the set has a *positive bounded spectrum* if $\nu_1^{\mathscr{M}} > 0$.

Assumption 4 *For all $\boldsymbol{x}' \in \mathscr{V}$, there exists a symmetric matrix $\boldsymbol{M}(\boldsymbol{x}')$ such that for all $\boldsymbol{x} \in \mathscr{V}$,*

$$Q(\boldsymbol{x},\boldsymbol{x}') = P(\boldsymbol{x}') + (\boldsymbol{x}-\boldsymbol{x}')^T \nabla P(\boldsymbol{x}') + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}')^T \boldsymbol{M}(\boldsymbol{x}')(\boldsymbol{x}-\boldsymbol{x}') \geqslant P(\boldsymbol{x}). \quad (11)$$

Moreover, the set $\{\boldsymbol{M}(\boldsymbol{x}) | \boldsymbol{x} \in \mathscr{V}\}$ has a nonnegative bounded spectrum with bounds $(\nu_1^{\mathscr{M}}, \nu_2^{\mathscr{M}})$.

As emphasized in [22, Lem.2.1], Assumption 4 is not a restrictive condition since it holds if P is gradient Lipschitz on $\mathscr{V}$ with constant L_p by setting $\boldsymbol{M}(\boldsymbol{x}) = L_p$ for all $\boldsymbol{x} \in \mathscr{V}$. Useful methods for constructing $M(\boldsymbol{x})$ without requiring the knowledge of L_p are developped in [5,18,13].

Assumption 5 *Assumption 4 holds and at least one of the following conditions is fulfilled:*

1) $\mathrm{Ker}(\boldsymbol{A}) = \{\boldsymbol{0}\}$ *with* $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]^T$
2) $\nu_1^{\mathscr{M}} > 0$.

Lemma 1 *If ψ_i is given by* (3), (4) *or* (5)*, then*

- *ψ_i is strictly convex*
- *$\dot{\psi}_i$ is strictly concave*
- $\lim_{u\to 0} \dot{\psi}_i(u) = -\infty$
- $-\dddot{\psi}_i(u)/\ddot{\psi}_i(u) \leqslant 2/u, \forall u > 0$

Proof In all cases, it is straightforward to check the first three conditions. The fourth also holds since we have:

1. $\psi_i(u) = -\kappa_i \log u,\ \kappa_i > 0 \implies -\dddot{\psi}_i(u)/\ddot{\psi}_i(u) = 2/u$
2. $\psi_i(u) = \kappa_i u \log u,\ \kappa_i > 0 \implies -\dddot{\psi}_i(u)/\ddot{\psi}_i(u) = 1/u \leqslant 2/u$
3. $\psi_i(u) = -\kappa_i u^r,\ r \in (0,1),\ \kappa_i > 0 \implies -\dddot{\psi}_i(u)/\ddot{\psi}_i(u) = (2-r)/u \leqslant 2/u$

□

3 Line search strategies for barrier functions

3.1 Problem statement

The stepsize should satisfy sufficient conditions to ensure the convergence of the descent algorithm. The most popular are the Wolfe conditions that state that a stepsize series $\{\alpha_k\}$ is acceptable if there exists c_1, $c_2 \in (0,1)$ such that for all k and for all $\boldsymbol{x}_k \in \mathscr{V}$,

$$F(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) \leqslant F(\boldsymbol{x}_k) + c_1 \alpha_k \boldsymbol{g}_k^T \boldsymbol{d}_k \quad (12)$$

$$|\nabla F(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k| \leqslant c_2 |\boldsymbol{g}_k^T \boldsymbol{d}_k| \quad (13)$$

where $\boldsymbol{g}_k \triangleq \nabla F(\boldsymbol{x}_k)$. The barrier term $B(\boldsymbol{x})$ implies that $\dot{f}$ tends to $-\infty$ when α is such that $C_i(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ cancels for some i. Since the constraints are linear, function f

is undefined outside an interval (α_-, α_+). Therefore, we must ensure that during the line search, the stepsize values remain in the interval (α_-, α_+).

Typical line search schemes in barrier-related optimization methods choose $\alpha_k = \theta\alpha_+$, where $\theta \in (0,1)$ is close to one ([34,15]). However, this simple approach does not ensure the convergence of the optimization algorithm and can lead to a sequence of iterates 'trapped' near the singularity ([27]). In [30,20], line search procedures based on the self-concordancy property of the logarithmic barrier functions are developped. However, the computation of the stepsize requires the evaluation of the Hessian matrix which is often expensive or even impossible for large scale problems. Furthermore, since methods using polynomial interpolation are not suited to interpolate function f, due to its behavior at α_- and α_+, [11,27] propose an interpolating function of the form

$$F(\boldsymbol{x} + \alpha \boldsymbol{d}) \approx f_0 + f_1\alpha + f_2\alpha^2 - \mu \log(f_3 - \alpha) \tag{14}$$

where the coefficients f_i are chosen to fit f and its derivative at two trial points. The line search strategy consists in repeating such a specific interpolation process until the fulfillment of Wolfe conditions. However, the resulting algorithm is not often used in practice, probably because the proposed interpolating function is difficult to compute. In contrast, our proposal is not based on interpolation, but rather on majorization, with a view to propose an analytical stepsize formula and to preserve strong convergence properties. Furthermore, the majorizing function and the resulting stepsize are easily computable.

3.2 Majoration-Minimization line search

In Majoration-Minimization (MM) algorithms ([18,19]), the minimization of a function f is obtained by performing successive minimizations of *tangent majorant* functions for f. Function $h(\boldsymbol{u}, \boldsymbol{v})$ is said tangent majorant for $f(\boldsymbol{u})$ at $\boldsymbol{v}$ if for all $\boldsymbol{u}$,

$$\begin{cases} h(\boldsymbol{u}, \boldsymbol{v}) \geqslant f(\boldsymbol{u}) \\ h(\boldsymbol{v}, \boldsymbol{v}) = f(\boldsymbol{v}) \end{cases}$$

The initial optimization problem is then replaced by a sequence of easier subproblems, corresponding to the MM update rule

$$\boldsymbol{u}^{j+1} = \arg\min_{\boldsymbol{u}} h(\boldsymbol{u}, \boldsymbol{u}^j).$$

Recently, the MM strategy has been used as a line search procedure ([12]) and the convergence is established in the case of conjugate-gradient ([37,22]), memory-gradient ([25]) and truncated Newton algorithms ([21]). The stepsize value α_k results from J successive minimizations of quadratic tangent majorant functions for the scalar function f, expressed as

$$q^j(\alpha, \alpha^j) = f(\alpha^j) + (\alpha - \alpha^j)\dot{f}(\alpha^j) + \frac{1}{2} m^j (\alpha - \alpha^j)^2 \tag{15}$$

at α^j. It is obtained by the recurrence

$$\alpha^0 = 0; \qquad \alpha^{j+1} = \alpha^j - \frac{\dot{f}(\alpha^j)}{m^j},\ j = 0,\ldots,J-1$$

and the stepsize α_k corresponds to the last value α^J. The main advantage of this procedure is that it gives an analytical formulation of the stepsize value and guarantees the algorithm convergence whatever the value of J ([22]). However, it cannot be applied in the case of logarithmic barrier function (3) since there is no parameter m^j such that the quadratic function $q^j(.,\alpha^j)$ majorizes f in the set (α_-,α_+). Actually, it would be sufficient to majorize f within the level set $\mathscr{L}_k = \{\alpha, F(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) \leqslant F(\boldsymbol{x}_k)\}$, but this set is difficult to determine or even to approximate. In the case of barriers (4) and (5), f is bounded at the boundary of the set (α_-,α_+). However, the curvature of f is unbounded and one can expect suboptimal results by majorizing the scalar function with a parabola. In particular, very small values of m^j will be obtained for α^j close to the singularity.

4 Proposed majorant function

To account for the barrier term, we propose the following form of tangent majorant function:

$$h(\alpha) = h_0 + h_1\alpha + h_2\alpha^2 - h_3 \log(h_4 - \alpha),$$

This form is reminiscent of the interpolation function (14) but here the parameters h_i are chosen to ensure the majorization property. Moreover, its minimizer can be calculated explicitely.

According to the MM theory, let us define the stepsize α_k by

$$\begin{aligned} \alpha^0 &= 0 \\ \alpha^{j+1} &= \arg\min_{\alpha} h^j(\alpha,\alpha^j), \quad j = 0,\ldots,J-1 \\ \alpha_k &= \alpha^J \end{aligned} \tag{16}$$

where $h^j(\alpha,\alpha^j)$ is the tangent majorant function

$$h^j(\alpha,\alpha^j) = q^j(\alpha,\alpha^j) + \gamma^j \left[(\bar{\alpha}^j - \alpha^j) \log\left(\frac{\bar{\alpha}^j - \alpha^j}{\bar{\alpha}^j - \alpha} \right) - \alpha + \alpha^j \right] \tag{17}$$

which depends on the value of f and its derivative at α^j and on three design parameters $m^j, \gamma^j, \bar{\alpha}^j$. It is easy to check that

$$h^j(\alpha^j,\alpha^j) = f(\alpha^j).$$

Thus, the values of $m^j, \gamma^j, \bar{\alpha}^j$ should ensure

$$h^j(\alpha,\alpha^j) \geqslant f(\alpha), \forall \alpha.$$

4.1 Construction of the majorant function

Let $\boldsymbol{x} \in \mathscr{C}$, $\boldsymbol{d}$ a search direction and $\alpha^j \in (\alpha_-, \alpha_+)$ such that $\boldsymbol{x} + \alpha^j \boldsymbol{d} \in \mathscr{V}$. Let us derive an expression for the parameters $m^j, \gamma^j, \bar{\alpha}^j$ such that $h^j(\alpha, \alpha^j)$ is a tangent majorant for $F(\boldsymbol{x} + \alpha \boldsymbol{d}) = f(\alpha)$ at α^j. Properties 1 and 2 respectively propose tangent majorant for $p(\alpha) \triangleq P(\boldsymbol{x} + \alpha \boldsymbol{d})$ and for $b(\alpha) \triangleq B(\boldsymbol{x} + \alpha \boldsymbol{d})$.

Property 1 Under Assumption 5, the function $q_p^j(\alpha, \alpha^j)$ given by $p(\alpha^j) + (\alpha - \alpha^j)\dot{p}(\alpha^j) + \frac{1}{2} m_p^j (\alpha - \alpha^j)^2$ is a tangent majorant for p at α^j if

$$m_p^j = \boldsymbol{d}^T M(\boldsymbol{x} + \alpha^j \boldsymbol{d}) \boldsymbol{d}. \tag{18}$$

Proof Direct consequence of Assumption 5. □

In order to build a tangent majorant for the barrier term b, we define

$$b_1(\alpha) = \sum_{i|\delta_i > 0} \psi_i(\theta_i + \alpha \delta_i)$$
$$b_2(\alpha) = \sum_{i|\delta_i < 0} \psi_i(\theta_i + \alpha \delta_i)$$

with $\theta_i = \boldsymbol{a}_i^T \boldsymbol{x} + \rho_i$ and $\delta_i = \boldsymbol{a}_i^T \boldsymbol{d}$ for all $i = 1, \ldots, m$ so that $b = b_1 + b_2 + \text{cste}$. Functions b_1 and b_2 present vertical asymptotes respectively at $\alpha_- < \alpha^j$ and $\alpha_+ > \alpha^j$ with

$$\begin{cases} \alpha_- = \max\limits_{i|\delta_i > 0} -\frac{\theta_i}{\delta_i}, \\ \alpha_+ = \min\limits_{i|\delta_i < 0} -\frac{\theta_i}{\delta_i}. \end{cases}$$

Property 2 The function $\phi^j(\alpha, \alpha^j)$ given by

$$b(\alpha^j) + (\alpha - \alpha^j)\dot{b}(\alpha^j) + \frac{1}{2} m_b^j (\alpha - \alpha^j)^2 + \gamma_b^j \left[(\bar{\alpha}^j - \alpha^j) \log \frac{\bar{\alpha}^j - \alpha^j}{\bar{\alpha}^j - \alpha} + \alpha^j - \alpha \right]$$

with parameters

$$m_b^j = \ddot{b}_1(\alpha^j), \quad \gamma_b^j = (\alpha_+ - \alpha^j)\ddot{b}_2(\alpha^j), \quad \bar{\alpha}^j = \alpha_+, \quad \text{for} \quad \alpha \in [\alpha^j, \alpha_+) \tag{19}$$

and

$$m_b^j = \ddot{b}_2(\alpha^j), \quad \gamma_b^j = (\alpha_- - \alpha^j)\ddot{b}_1(\alpha^j), \quad \bar{\alpha}^j = \alpha_-, \quad \text{for} \quad \alpha \in (\alpha_-, \alpha^j] \tag{20}$$

is a tangent majorant for b at α^j.

Proof Let us first prove this property for $\alpha \geqslant \alpha^j$. In this case, function ϕ^j is noted ϕ^j_+ with parameters $m^j_+ = m^j_b$ and $\gamma^j_+ = \gamma^j_b$. The aim is to prove that

$$\begin{cases} \phi^j_{+1}(\alpha,\alpha^j) = b_1(\alpha^j) + (\alpha-\alpha^j)\dot{b}_1(\alpha^j) + \frac{1}{2}m^j_+(\alpha-\alpha^j)^2 \\ \phi^j_{+2}(\alpha,\alpha^j) = b_2(\alpha^j) + (\alpha-\alpha^j)\dot{b}_2(\alpha^j) + \gamma^j_+\left[(\alpha_+-\alpha^j)\log\frac{\alpha_+-\alpha^j}{\alpha_+-\alpha} + \alpha^j - \alpha\right] \end{cases}$$

respectively majorize b_1 and b_2 for all $\alpha \geqslant \alpha_j$.

First, Lemma 1 implies that b_1 is strictly convex and $\dot{b}_1$ is strictly concave. Then, for all $\alpha \in [\alpha_j;\alpha^+)$, $\ddot{b}_1(\alpha) \leqslant \ddot{b}_1(\alpha^j) = m^j_+$. Hence, $\phi^j_{+1}(.,\alpha^j)$ majorizes b_1 on $[\alpha_j;\alpha^+)$.

Then, let us define $T(\alpha) = \dot{b}_2(\alpha)(\alpha_+-\alpha)$ and $l(\alpha) = \dot{b}_2(\alpha^j)(\alpha_+-\alpha) + \gamma^j_+(\alpha-\alpha^j)$. Given $\gamma^j_+ = (\alpha_+-\alpha^j)\ddot{b}_2(\alpha^j)$, the linear function l also reads:

$$l(\alpha) = \dot{\phi}^j_{+2}(\alpha,\alpha^j)(\alpha_+-\alpha)$$

Thus we have $l(\alpha^j) = T(\alpha^j)$ and $\dot{l}(\alpha^j) = \dot{T}(\alpha^j)$. Moreover:

$$\ddot{T}(\alpha) = \dddot{b}_2(\alpha)(\alpha_+-\alpha) - 2\ddot{b}_1(\alpha) \quad = \sum_{i|\delta_i<0} \delta_i^3\dddot{\psi}_i(\theta_i+\alpha\delta_i)(\alpha_+-\alpha) - 2\delta_i^2\ddot{\psi}_i(\theta_i+\alpha\delta_i) \tag{21}$$

According to the definition of α_+:

$$(\alpha_+-\alpha) < -(\theta_i+\alpha\delta_i)/\delta_i, \forall i \,\text{such that}\, \delta_i < 0$$

According to Lemma 1, the third derivative of ψ_i is negative, so

$$\ddot{T}(\alpha) < \sum_{i|\delta_i<0} \delta_i^2\left[-\dddot{\psi}_i(\theta_i+\alpha\delta_i)(\theta_i+\alpha\delta_i) - 2\ddot{\psi}_i(\theta_i+\alpha\delta_i)\right] < 0$$

where the last inequality is a consequence of Lemma 1. Thus T is concave. Since l is a linear function tangent to T, we have

$$l(\alpha) \geqslant T(\alpha), \forall\alpha \in [\alpha_j,\alpha^+) \tag{22}$$

Given $\alpha_+ > \alpha$, (22) also reads:

$$\dot{\phi}^j_{+2}(\alpha,\alpha^j) \geqslant \dot{b}_2(\alpha), \forall\alpha \in [\alpha_j,\alpha^+) \tag{23}$$

Therefore, $\phi^j_{+2}(.,\alpha^j)$ majorizes b_2 over $[\alpha_j;\alpha^+)$. Finally, $\phi^j_+(.,\alpha^j) = \phi^j_{+1}(.,\alpha^j) + \phi^j_{+2}(.,\alpha^j)$ majorizes b for $\alpha \geqslant \alpha_j$.

The same elements of proof apply to the case $\alpha \leqslant \alpha^j$. □

Therefore, using Properties 1 and 2, we obtain that $h^j(\alpha,\alpha^j) = q^j_p(\alpha,\alpha^j) + \mu\phi^j(\alpha,\alpha^j)$ is a tangent majorant for f at α^j.

4.2 Minimization of the tangent majorant

The MM recurrence (16) involves the computation of the minimizer of $h^j(\alpha,\alpha^j)$ for $j \in \{0,\dots,J-1\}$. Lemma 2 leads to the strict convexity of the tangent majorant:

Lemma 2 *Under Assumption 5, $h^j(.,\alpha^j)$ is C^2 and strictly convex.*

Proof First, $q_p^j(.,\alpha^j)$ is a quadratic function and thus C^2 over (α_-,α_+). Moreover, $h^j(.,\alpha^j)$ is C^∞ over $(\alpha_-;\alpha^j)$ and $(\alpha^j;\alpha_+)$. Finally, expressions (19) and (20) lead to the continuity of h^j and of its first and second derivatives at α^j. Then, $h^j(.,\alpha^j)$ is C^2 over $(\alpha_-;\alpha_+)$. According to (19) and (20), the second derivative of $h^j(.,\alpha^j)$ is given by

$$\ddot{h}^j(\alpha,\alpha^j) = \begin{cases} m_p^j + \mu\ddot{b}_2(\alpha^j) + \mu\ddot{b}_1(\alpha^j)\frac{(\alpha_- - \alpha^j)^2}{(\alpha_- - \alpha)^2} & \forall\alpha \in (\alpha_-,\alpha^j] \\ m_p^j + \mu\ddot{b}_1(\alpha^j) + \mu\ddot{b}_2(\alpha^j)\frac{(\alpha_+ - \alpha^j)^2}{(\alpha_+ - \alpha)^2} & \forall\alpha \in [\alpha^j,\alpha_+) \end{cases}$$

m_p^j is strictly positive according to Assumption 5, and b_1 and b_2 are strictly convex according to Lemma 1. Hence, $h^j(.,\alpha^j)$ is strictly convex. □

Because of strict convexity, the tangent majorant $h^j(.,\alpha^j)$ has a unique minimizer, which can be expressed as an explicit function of $\dot{f}(\alpha^j)$ as follows:

$$\alpha^{j+1} = \begin{cases} \alpha^j - \dfrac{2q_3}{q_2 + \sqrt{q_2^2 - 4q_1q_3}} & \text{if} \quad \dot{f}(\alpha^j) \leqslant 0 \\ \alpha^j - \dfrac{2q_3}{q_2 - \sqrt{q_2^2 - 4q_1q_3}} & \text{if} \quad \dot{f}(\alpha^j) > 0 \end{cases} \tag{24}$$

with

$$\begin{cases} q_1 = -m^j \\ q_2 = \gamma^j - \dot{f}(\alpha^j) + m^j(\bar{\alpha}^j - \alpha^j) \\ q_3 = (\bar{\alpha}^j - \alpha^j)\dot{f}(\alpha^j) \end{cases} \tag{25}$$

4.3 Properties of the tangent majorant

Lemma 3 *Let $j \in \{0,\dots,J-1\}$. If $\dot{f}(\alpha^j) \leqslant 0$, then α^{j+1} fulfills:*

$$-\frac{q_3}{q_2} \leqslant \alpha^{j+1} - \alpha^j \leqslant -\frac{2q_3}{q_2}.$$

where q_1, q_2 and q_3 are given by (25).

Proof Straightforward given (24) with $\dot{f}(\alpha^j) \leqslant 0$. □

Lemma 4 *Let $j \in \{0,\dots,J-1\}$. For all $\alpha \in [\alpha^j,\alpha_+)$, $\dot{\phi}_+^j(\alpha,\alpha^j)$ majorizes the derivative $\dot{b}(\alpha)$.*

Proof For all α^j, we have

$$\ddot{\phi}^j_{+1}(\alpha, \alpha^j) = \ddot{b}_1(\alpha^j) \geqslant \ddot{b}_1(\alpha), \forall \alpha \in [\alpha^j, \alpha_+)$$

Thus, function $\dot{\phi}^j_{+1}(\alpha, \alpha^j) - \dot{b}_1(\alpha)$ is increasing on $[\alpha^j; \alpha_+)$. Moreover, it vanishes at α^j, so

$$\dot{\phi}^j_{+1}(\alpha, \alpha^j) \geqslant \dot{b}_1(\alpha), \forall \alpha \in [\alpha^j, \alpha_+)$$

This allows to conclude, given (23). □

Property 3 Let $j \in \{0, \ldots, J-1\}$. Under Assumptions 1 and 5, there exists $\nu_{\min}$, $\nu_{\max}$, $0 < \nu_{\min} \leqslant \nu_{\max}$, such that for all $\boldsymbol{x} \in \mathscr{V}$ and for all descent direction $\boldsymbol{d}$ at $\boldsymbol{x}$:

$$\nu_{\min} \|\boldsymbol{d}\|^2 \leqslant \ddot{h}^j(\alpha^j, \alpha^j) \leqslant \nu_{\max} \|\boldsymbol{d}\|^2, \forall j \geqslant 0$$

Proof According to Lemma 2,

$$\ddot{h}^j(\alpha^j, \alpha^j) = m^j_p + \mu \ddot{b}(\alpha^j).$$

The second derivative of b at α^j also reads

$$\ddot{b}(\alpha^j) = \boldsymbol{d}^T \nabla^2 B(\boldsymbol{x} + \alpha^j \boldsymbol{d}) \boldsymbol{d}$$

and Property 1 gives

$$m^j_p = \boldsymbol{d}^T M(\boldsymbol{x} + \alpha^j \boldsymbol{d}) \boldsymbol{d}.$$

Moreover, $\boldsymbol{x} + \alpha^j \boldsymbol{d} \in \mathscr{V}$. Thus, it is sufficient to show that the set $\left\{M(\boldsymbol{x}) + \mu \nabla^2 B(\boldsymbol{x}) | \boldsymbol{x} \in \mathscr{V}\right\}$ has a positive bounded spectrum.Let $\boldsymbol{x} \in \mathscr{V}$.

$$\nabla^2 B(\boldsymbol{x}) = \boldsymbol{A}^T \operatorname{diag}(\tau_i C_i(\boldsymbol{x})^{-t_i}) \boldsymbol{A} \tag{26}$$

with

$$(\tau_i, t_i) = \begin{cases} (2, \kappa_i) & \text{if} \quad \phi_i(u) = -\kappa_i \log u \\ (1, 1) & \text{if} \quad \phi_i(u) = u \log u \\ (-r^2 + r, 2 - r) & \text{if} \quad \phi_i = -u^r \end{cases}$$

and $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]^T$. $\boldsymbol{x} \in \mathscr{V}$ so (9) and (8) yield

$$\boldsymbol{d}^T \boldsymbol{T}(M) \boldsymbol{d} \leqslant \boldsymbol{d}^T \nabla^2 B(\boldsymbol{x}) \boldsymbol{d} \leqslant \boldsymbol{d}^T \boldsymbol{T}(\varepsilon_0) \boldsymbol{d} \tag{27}$$

with $\boldsymbol{T}(m) = \boldsymbol{A}^T \operatorname{diag}(\tau_i m^{-t_i}) \boldsymbol{A}$. Matrix $\boldsymbol{T}(m)$ is symmetric and has a nonnegative bounded spectrum with bounds $(\nu^{\mathscr{T}}_{\min}(m), \nu^{\mathscr{T}}_{\max}(m))$. Moreover, according to Assumption 4, $M(\boldsymbol{x})$ has a nonnegative bounded spectrum with bounds $(\nu^{\mathscr{M}}_{\min}, \nu^{\mathscr{M}}_{\max})$. Finally, according to Assumption 5, either $\nu^{\mathscr{M}}_{\min} > 0$ or $\operatorname{Ker}(\boldsymbol{A}^T \boldsymbol{A}) = \{\boldsymbol{0}\}$. Since the latter condition implies $\nu^{\mathscr{T}}_{\min}(m) > 0$, Property 3 holds with $\nu_{\min} = \nu^{\mathscr{M}}_{\min} + \mu \nu^{\mathscr{T}}_{\min}(M) > 0$ and $\nu_{\max} = \nu^{\mathscr{M}}_{\max} + \mu \nu^{\mathscr{T}}_{\max}(\varepsilon_0)$. □

5 Properties of the stepsize series

This section presents essential properties of the stepsize series (16) allowing to establish the convergence conditions of the descent algorithm. Let us consider $\boldsymbol{x} \in \mathscr{V}$ and a descent direction $\boldsymbol{d}$, so that $\dot{f}(0) = \boldsymbol{d}^T \boldsymbol{g} < 0$. The MM recurrence produces monotonically decreasing values $\{f(\alpha^j)\}$ and the series $\{\alpha^j\}$ converges to a stationnary point of f ([18]). Moreover, it is readily seen from (24) that

$$\operatorname{sgn}\left(\alpha^{j+1} - \alpha^j\right) = -\operatorname{sgn}\left(\dot{f}(\alpha^j)\right), \forall j \geqslant 0 \tag{28}$$

Furthermore, according to [19, Th.6.4], the set $[0, \tilde{\alpha}]$ with $\tilde{\alpha} = \min\{\alpha > 0 | f(\alpha) = f(0)\}$ acts as a capture basin, that is

$$\alpha^j \in [0, \tilde{\alpha}], \quad \forall j \geqslant 0. \tag{29}$$

Since $\dot{f}(0) < 0$, it can easily been shown that α^1 is strictly positive so $\alpha^j > 0$ for all $j \geqslant 1$ using the capture property (29). We have finally the following result:

Lemma 5 *If for all* $j \in \{0, \ldots, J-1\}$,

$$\dot{p}(\alpha) \leqslant \dot{q}_p^j(\alpha, \alpha^j), \forall \alpha \geqslant \alpha^j \tag{30}$$

then

$$\dot{f}(\alpha^j) \leqslant 0, \forall j \in \{0, \ldots, J-1\} \tag{31}$$

and the series $\left\{\alpha^j\right\}$ *is nondecreasing.*

Proof According to Lemma 5, (30) implies that for all $j \in \{0, \ldots, J-1\}$,

$$\dot{f}(\alpha) \leqslant \dot{h}^j(\alpha, \alpha^j), \forall \alpha \geqslant \alpha^j. \tag{32}$$

Moreover, (31) holds for $j = 0$ since $\boldsymbol{d}$ is a descent direction. Thus, $\alpha^1 \geqslant 0$ according to (28). Let $j \in \{0, \ldots, J-1\}$ and assume that $\dot{f}(\alpha^j) \leqslant 0$. Thus, according to (28), $\alpha^{j+1} \geqslant \alpha^j$. Using (32) for $\alpha = \alpha^{j+1}$, we obtain:

$$\dot{f}(\alpha^{j+1}) \leqslant \dot{h}^j(\alpha^{j+1}, \alpha^j)$$

Moreover α^{j+1} is the minimizer of $h^j(., \alpha^j)$ so $\dot{h}^j(\alpha^{j+1}, \alpha^j) = 0$, hence the result by immediate recurrence on j. □

5.1 Lower and upper bounds for the stepsize

Property 4 Under Assumptions 1 and 5, there exist $\nu, \nu' > 0$ such that

$$\frac{-\boldsymbol{g}^T \boldsymbol{d}}{\nu \|\boldsymbol{d}\|^2} \leqslant \alpha^1 \leqslant \frac{-\boldsymbol{g}^T \boldsymbol{d}}{\nu' \|\boldsymbol{d}\|^2}. \tag{33}$$

Proof $\boldsymbol{d}$ is a descent direction, so $\dot{f}(0) < 0$ and $h^0(.,0)$ has a barrier at $\overline{\alpha}^0 = \alpha_+$.

If $\alpha_+ = +\infty$ then $h^0(.,0)$ is a quadratic function with curvature m^0. This majorant is minimized at $\alpha^1 = -\dot{f}(0)/m^0$ and according to Property 3, we have:

$$\frac{-\boldsymbol{g}^T\boldsymbol{d}}{\nu_{\max}\|\boldsymbol{d}\|^2} \leqslant \alpha^1 \leqslant \frac{-\boldsymbol{g}^T\boldsymbol{d}}{\nu_{\min}\|\boldsymbol{d}\|^2}$$

If $\alpha_+ < +\infty$, according to Lemma 3:

$$\frac{-\boldsymbol{g}^T\boldsymbol{d}}{\frac{\gamma^0}{\alpha_+} - \frac{\boldsymbol{g}^T\boldsymbol{d}}{\alpha_+} + m^0} \leqslant \alpha^1 \leqslant \frac{-2\boldsymbol{g}^T\boldsymbol{d}}{\frac{\gamma^0}{\alpha_+} - \frac{\boldsymbol{g}^T\boldsymbol{d}}{\alpha_+} + m^0}$$

Using Property 3 and the positivity of $-\boldsymbol{g}^T\boldsymbol{d}$, we obtain

$$\nu_{\min}\|\boldsymbol{d}\|^2 \leqslant \frac{\gamma^0}{\alpha_+} - \frac{\boldsymbol{g}^T\boldsymbol{d}}{\alpha_+} + m^0 \tag{34}$$

On the other hand, taking $\iota = \arg\max_i -\boldsymbol{a}_i^T\boldsymbol{d}$, we deduce from (8) that

$$\alpha^+ \geqslant \frac{\varepsilon_0}{|\boldsymbol{a}_\iota^T\boldsymbol{d}|}.$$

Thus, using Cauchy-Schwartz inequality and (7),

$$\begin{aligned}\frac{-\boldsymbol{g}^T\boldsymbol{d}}{\alpha_+} = \frac{|\boldsymbol{g}^T\boldsymbol{d}|}{\alpha_+} &\leqslant |\boldsymbol{g}^T\boldsymbol{d}|.|\boldsymbol{a}_\iota^T\boldsymbol{d}|\frac{1}{\varepsilon_0} \\ &\leqslant \|\boldsymbol{g}\|\|\boldsymbol{a}_\iota\|\|\boldsymbol{d}\|^2\frac{1}{\varepsilon_0} \\ &\leqslant \frac{\eta\mathscr{A}}{\varepsilon_0}\|\boldsymbol{d}\|^2 \end{aligned} \tag{35}$$

with $\mathscr{A} = \max_i \|\boldsymbol{a}_i\| > 0$. Moreover, Property 3 implies that there exists $\nu_{\max}$ such that

$$m^0 + \frac{\gamma^0}{\alpha_+} \leqslant \nu_{\max}\|\boldsymbol{d}\|^2 \tag{36}$$

Therefore (34), (35) and (36) allow to check that Property 4 holds for

$$\begin{aligned}\nu &= \nu_{\max} + \eta\mathscr{A}/\varepsilon_0 \\ \nu' &= \nu_{\min}/2\end{aligned}$$

$\square$

5.2 Sufficient decrease condition

The first Wolfe condition (12) measures whether the stepsize value induces a sufficient decrease of F. It also reads

$$f(\alpha) - f(0) \leqslant c_1 \alpha \dot{f}(0). \tag{37}$$

where $c_1 \in (0,1)$ is a constant with respect to the iteration number.

In this section, we show that (37) holds with the stepsize value produced by the proposed MM strategy. First, we need the following lemmas.

Lemma 6 *Let $j \in \{0,\dots,J-1\}$. If $\dot{f}(\alpha^j) \leqslant 0$, then:*

$$f(\alpha^j) - f(\alpha^{j+1}) + \frac{1}{2}(\alpha^{j+1} - \alpha^j)\dot{f}(\alpha^j) \geqslant 0 \tag{38}$$

Proof The property is trivial if $\dot{f}(\alpha^j) = 0$. Assume that $\dot{f}(\alpha^j) < 0$ so that $\alpha_+ > \alpha^{j+1} > \alpha^j$. Let define the function $\xi : u \to -\log(1-u) - u$. A straightforward analysis of ξ shows that

$$\frac{\xi(u)}{u\dot{\xi}(u)} \leqslant \frac{1}{2}, \quad \forall u \in (0,1) \tag{39}$$

Taking $u = \frac{\alpha-\alpha^j}{\alpha_+-\alpha^j}$ in (39) and denoting $\varphi(\alpha) = \xi(u)$:

$$\frac{\varphi(\alpha)}{(\alpha-\alpha^j)\dot{\varphi}(\alpha)} \leqslant \frac{1}{2}, \forall \alpha \in (\alpha^j; \alpha_+). \tag{40}$$

Moreover, let us define $Q(\alpha) = \frac{1}{2}m^j(\alpha-\alpha^j)^2$ so that

$$Q(\alpha) = \frac{1}{2}(\alpha-\alpha^j)\dot{Q}(\alpha). \tag{41}$$

Let $\tau(\alpha) = Q(\alpha) + \gamma^j(\alpha_+ - \alpha^j)\varphi(\alpha)$ so the majorant function reads

$$h^j(\alpha,\alpha^j) = f(\alpha^j) + (\alpha-\alpha^j)\dot{f}(\alpha^j) + \tau(\alpha), \forall \alpha \in [\alpha^j, \alpha_+)$$

and, using (40) and (41),

$$\frac{\tau(\alpha)}{(\alpha-\alpha^j)\dot{\tau}(\alpha)} \leqslant \frac{1}{2}, \forall \alpha \in (\alpha^j; \alpha_+) \tag{42}$$

$h^j(.,\alpha^j)$ is a tangent majorant for f so

$$h^j(\alpha,\alpha^j) - f(\alpha) = f(\alpha^j) - f(\alpha) + (\alpha-\alpha^j)\dot{f}(\alpha^j) + \tau(\alpha) \geqslant 0 \tag{43}$$

Taking $\alpha = \alpha^{j+1} > \alpha^j$ in (42) and (43), we obtain

$$f(\alpha^j) - f(\alpha^{j+1}) + (\alpha^{j+1} - \alpha^j)\dot{f}(\alpha^j) + \frac{1}{2}(\alpha^{j+1} - \alpha^j)\dot{\tau}(\alpha^{j+1}) \geqslant 0$$

Hence the result using

$$\begin{aligned}\dot{\tau}(\alpha^{j+1}) &= \dot{h}^j(\alpha^{j+1},\alpha^j) - \dot{f}(\alpha^j)\\ &= -\dot{f}(\alpha^j)\end{aligned}$$

□

Lemma 7 *Under Assumptions 1 and 5, for all* $j \in \{1,\dots,J\}$,

$$\alpha^j \leqslant c^j_{\max}\alpha^1, \tag{44}$$

where

$$c^j_{\max} = \left(1+\frac{2\nu_{\max}L}{\nu^2_{\min}}\right)^{j-1}\left(1+\frac{\nu}{L}\right)-\frac{\nu}{L} \geqslant 1. \tag{45}$$

Proof It is easy to check (44) for $j=1$, with $c^1_{\max}=1$. Let us prove that (44) holds for $j>1$. Assume that $\dot{f}(\alpha^j)<0$. Then $\bar{\alpha}^j=\alpha_+$ and we can deduce from Lemma 3 that

$$\begin{aligned}\alpha^{j+1}-\alpha^j &\leqslant \frac{-2\dot{f}(\alpha^j)}{(\gamma^j-\dot{f}(\alpha^j))/(\alpha_+-\alpha^j)+m^j}\\ &\leqslant \frac{-2\dot{f}(\alpha^j)}{\gamma^j/(\alpha_+-\alpha^j)+m^j}\end{aligned} \tag{46}$$

According to Property 3:

$$\|\boldsymbol{d}\|^2 \geqslant \left(\gamma^0/\alpha_+ + m^0\right)/\nu_{\max} \tag{47}$$

and

$$\gamma^j/(\alpha_+-\alpha^j)+m^j \geqslant \nu_{\min}\|\boldsymbol{d}\|^2$$

thus we have

$$\gamma^j/(\alpha_+-\alpha^j)+m^j \geqslant \nu_{\min}\left(\gamma^0/\alpha_+ + m^0\right)/\nu_{\max} > 0$$

Then, from (46):

$$\alpha^{j+1} \leqslant \alpha^j + |\dot{f}(\alpha^j)|\frac{2\nu_{\max}}{(\gamma^0/\alpha_+ + m^0)\nu_{\min}} \tag{48}$$

If $\dot{f}(\alpha^j)\geqslant 0$, α^{j+1} is lower than α^j so (48) still holds. According to Assumption 1, ∇F is Lipschitz, so that:

$$|\dot{f}(\alpha^j)-\dot{f}(0)| \leqslant L\|\boldsymbol{d}\|^2\alpha^j$$

Using the fact that $|\dot{f}(\alpha^j)| \leqslant |\dot{f}(\alpha^j)-\dot{f}(0)|+|\dot{f}(0)|$, and $\dot{f}(0)<0$, we get:

$$|\dot{f}(\alpha^j)| \leqslant L\alpha^j\|\boldsymbol{d}\|^2-\dot{f}(0) \tag{49}$$

Using Property 4 and (47):

$$\begin{aligned}-\dot{f}(0) &\leqslant \alpha^1\nu\|\boldsymbol{d}\|^2\\ &\leqslant \alpha^1\nu(m^0+\gamma_0/\alpha_+)/\nu_{\min}\end{aligned} \tag{50}$$

Given (47)- (50), we get:

$$\alpha^{j+1} \leqslant \alpha^j + \frac{2\nu_{\max}}{(m^0+\gamma_0/\alpha_+)\nu_{\min}}\left[L\alpha^j\left(\frac{m^0+\gamma_0/\alpha_+}{\nu_{\min}}\right)+\alpha^1\frac{\nu}{\nu_{\min}}(m^0+\gamma_0/\alpha_+)\right]$$

Hence

$$\alpha^{j+1} \leqslant \alpha^j \left(1 + \frac{2\nu_{\max} L}{\nu_{\min}^2}\right) + 2\alpha^1 \frac{\nu_{\max}\nu}{\nu_{\min}^2}$$

This corresponds to a recursive definition of the series $(c_{\max}^j)$ with:

$$c_{\max}^{j+1} = c_{\max}^j \left(1 + 2\frac{\nu_{\max} L}{\nu_{\min}^2}\right) + 2\frac{\nu\nu_{\max}}{\nu_{\min}^2}$$

Given $c_{\max}^1 = 1$, (45) is the general term of the series. □

Property 5 (first Wolfe condition) Under Assumptions 1 and 5, the iterates of (16) fulfill

$$f(\alpha^j) - f(0) \leqslant c_1^j \alpha^j \dot{f}(0) \tag{51}$$

for all $j \geqslant 1$, with $c_1^j = (2c_{\max}^j)^{-1} \in (0,1)$.

Proof For $j = 1$, (51) holds according to Lemma 6, since it identifies with (38) when $j = 0$, given $c_{\max}^1 = 1$. For all $j > 1$, (51) holds by immediate recurrence, given Lemma 7. □

Property 5 corresponds to a strong result related to the proposed MM line search since it implies that the computed stepsize leads to a sufficient decrease of the criterion at each iteration, independently from the number of line search iterates J.

5.3 Stepsize minoration

Condition (12) alone is not sufficient to ensure that the algorithm makes reasonable progress since it holds for arbitrary small values for α and thus can yield convergence to a non-stationnary point ([31]). In order to avoid too short steps, a second condition is required, for example the second Wolfe condition (13). It turned out difficult or even impossible to establish the curvature condition (13) for any value of J. Fortunately, we can obtain a direct minoration of the stepsize values that is sufficient to yield convergence results.

Property 6 Under Assumptions 1 and 5, for all $j \geqslant 1$,

$$\alpha^j \geqslant c_{\min} \alpha^1 \tag{52}$$

and

$$\alpha^j \geqslant c_{\min} \frac{-\boldsymbol{g}^T \boldsymbol{d}}{\nu \|\boldsymbol{d}\|^2} \tag{53}$$

for some $c_{\min} > 0$.

Proof First, let us show that (52) holds for all $j \geqslant 1$ with

$$c_{\min} = \frac{\sqrt{1+2L/v_{\min}}-1}{2L/v_{\min}} \in (0,1/2) \tag{54}$$

Let ϕ be the concave quadratic function:

$$\phi(\alpha) = f(0) + \alpha \dot{f}(0) + m\frac{\alpha^2}{2}$$

with $m = -L(m^0 + \gamma^0/\alpha_+)/v_{\min}$. We have $\phi(0) = f(0)$ and $\dot{\phi}(0) = \dot{f}(0) < 0$, so ϕ is decreasing on $\mathbb{R}^+$. Let us consider $\alpha \in [0, \alpha^j]$, so that $\boldsymbol{x} + \alpha \boldsymbol{d} \in \mathscr{V}$. According to Assumption 1, we have

$$|\dot{f}(\alpha) - \dot{f}(0)| \leqslant \|\boldsymbol{d}\|^2 L |\alpha|$$

and according to Property 3,

$$|\dot{f}(\alpha) - \dot{f}(0)| \leqslant L\alpha(m^0 + \gamma^0/\alpha_+)/v_{\min}$$

Then we obtain:

$$|\dot{f}(\alpha)| \leqslant L\alpha(m^0 + \gamma^0/\alpha_+)/v_{\min} - \dot{f}(0)$$

Hence:

$$\dot{\phi}(\alpha) \leqslant \dot{f}(\alpha), \quad \forall \alpha \in [0, \alpha^j] \tag{55}$$

Integrating (55) between 0 and α^j yields

$$\phi(\alpha^j) \leqslant f(\alpha^j) \tag{56}$$

On the other hand, the expression of ϕ at $\alpha_{\min} = c_{\min}\alpha^1$ can be written as follows:

$$\phi(\alpha_{\min}) = f(0) + C\alpha^1 \dot{f}(0)$$

where

$$C = c_{\min} - c_{\min}^2 L\alpha^1 \frac{m^0 + \gamma^0/\alpha_+}{2\dot{f}(0)v_{\min}}.$$

According to (46):

$$\alpha^1 \leqslant \frac{-2\dot{f}(0)}{m^0 + \gamma^0/\alpha_+},$$

so that

$$C \leqslant c_{\min} + c_{\min}^2 \frac{L}{v_{\min}} = \frac{1}{2},$$

where the latter equality directly stems from the expression of $c_{\min}$. Since ϕ is decreasing on $\mathbb{R}^+$, we get

$$\phi(\alpha_{\min}) \geqslant f(0) + \frac{1}{2}\alpha^1 \dot{f}(0) \geqslant f(\alpha^1), \tag{57}$$

where the last inequality is the first Wolfe condition (51) for $j = 1$.

Finally, $\alpha^j > 0$ for all $j \geqslant 1$. Assume that there exists j such that $\alpha^j < \alpha_{\min}$. According to (56) and given that ϕ is decreasing on $\mathbb{R}^+$, we get:

$$f(\alpha^j) \geqslant \phi(\alpha^j) > \phi(\alpha_{\min}) \geqslant f(\alpha^1),$$

which contradicts the fact that $f(\alpha^j)$ is nonincreasing. Thus, (52) holds. So does (53), according to Property 4. □

6 Convergence results

This section discusses the convergence of the iterative descent algorithm

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \quad k = 1, \ldots, K$$

when $\boldsymbol{d}_k$ satisfies $\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$ and the line search is performed using the proposed MM strategy.

6.1 Zoutendijk condition

The global convergence of a descent direction method is not only ensured by a 'good choice' of the step but also by well-chosen search directions $\boldsymbol{d}_k$. Convergence proofs often rely on the fulfillment of Zoutendijk condition

$$\sum_{k=0}^{\infty} \|\boldsymbol{g}_k\|^2 \cos^2 \theta_k < \infty, \tag{58}$$

where θ_k is the angle between $\boldsymbol{d}_k$ and the steepest descent direction $-\boldsymbol{g}_k$:

$$\cos \theta_k = \frac{-\boldsymbol{g}_k^T \boldsymbol{d}_k}{\|\boldsymbol{g}_k\| \, \|\boldsymbol{d}_k\|}.$$

Inequality (58) implies that $\|\boldsymbol{g}_k\| \cos \theta_k$ vanishes for large values of k. Moreover, provided that $\boldsymbol{d}_k$ is not orthogonal to $-\boldsymbol{g}_k$ (i.e., $\cos \theta_k > 0$), condition (58) implies the convergence of the algorithm in the sense

$$\lim_{k \to \infty} \|\boldsymbol{g}_k\| = 0. \tag{59}$$

Zoutendijk condition holds when the line search procedure is based on the fulfillment of the sufficient conditions (12),(13) ([31]). In the case of the proposed line search, the following result holds.

Property 7 Let α_k be defined by (16). Under Assumptions 1 and 5, Zoutendijk condition (58) holds.

Proof Let us first remark that for all k, $\boldsymbol{d}_k \neq 0$, since $\boldsymbol{g}_k^T \boldsymbol{d}_k < 0$. According to Property 5, the first Wolfe condition holds for $c_1 = c_1^J$:

$$F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1}) \geqslant -c_1^J \alpha_k \boldsymbol{g}_k^T \boldsymbol{d}_k$$

According to Property 6:

$$\alpha_k \geqslant c_{\min} \frac{-\boldsymbol{g}_k^T \boldsymbol{d}_k}{\nu \|\boldsymbol{d}_k\|^2}$$

Hence:

$$0 \leqslant c_0 \frac{(\boldsymbol{g}_k^T \boldsymbol{d}_k)^2}{\|\boldsymbol{d}_k\|^2} \leqslant F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1})$$

with $c_0 = (c_{\min} c_1^J)/\nu > 0$. According to Assumption 1, the level set $\mathscr{L}_0$ is bounded, so $\lim_{k\to\infty} F(\boldsymbol{x}_k)$ is finite. Therefore:

$$\sum_{k=0}^{\infty} \frac{(\boldsymbol{g}_k^T \boldsymbol{d}_k)^2}{\|\boldsymbol{d}_k\|^2} \leqslant \frac{1}{c_0} \left[\lim_{k\to\infty} F(\boldsymbol{x}_k) - F(\boldsymbol{x}_0) \right] < \infty \tag{60}$$

□

6.2 Gradient related algorithms

A general convergence result can be established by using the concept of *gradient related* direction ([1]).

Definition 2 A direction sequence $\{\boldsymbol{d}_k\}$ is said gradient related to $\{\boldsymbol{x}_k\}$ if the following property holds: for any subsequence $\{\boldsymbol{x}_k\}_{\mathscr{K}}$ that converges to a nonstationary point, the corresponding subsequence $\{\boldsymbol{d}_k\}_{\mathscr{K}}$ is bounded and satisfies

$$\limsup_{k\to\infty, k\in\mathscr{K}} \boldsymbol{g}_k^T \boldsymbol{d}_k < 0.$$

Theorem 1 *([35]) Let $\{\boldsymbol{x}_k\}$ a sequence generated by a descent method $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$. Assume that the sequence $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$ and that Zoutendijk condition* (58) *holds. Then, the descent algorithm converges in the sense* $\lim_{k\to\infty} \|\boldsymbol{g}_k\| = 0$.

The gradient norm converging to zero does not imply that the optimization method converges to a minimizer, but only that it is attracted by a stationary point. However, under certain sufficient conditions, this can guarantee convergence to a local or global minimum.

Corollary 1 *Let $\{\boldsymbol{x}_k\}$ a sequence generated by a descent method $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$. Assume that the sequence $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$ and that Zoutendijk condition* (58) *holds. If $\lim_{k\to\infty} \nabla^2 F(\boldsymbol{x}_k)$ is positive definite then $\{\boldsymbol{x}_k\}$ converges to a strict local minimizer of F.*

Proof Direct consequence of the sufficient condition for local minimization ([31]).

Corollary 2 *Let $\{\boldsymbol{x}_k\}$ a sequence generated by a descent method $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$. Assume that the sequence $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$ and that Zoutendijk condition* (58) *holds. If Assumption 2 holds then $\{\boldsymbol{x}_k\}$ converges to a global minimizer of F.*

Proof Direct consequence of the sufficient condition for global minimization ([31]).

In the sequel, we will show that Theorem 1 yields convergence of classical descent optimization schemes such as the truncated Newton method and the projected gradient method for constrained optimization when such schemes are combined with our line search strategy.

6.2.1 Preconditioned gradient, Newton and inexact Newton algorithms

Let us consider the family of descent algorithms when the search direction has the form

$$\boldsymbol{d}_k = -\boldsymbol{D}_k \boldsymbol{g}_k \tag{61}$$

with $\boldsymbol{D}_k$ a symmetric and positive definite (SPD) matrix. In the steepest descent method $\boldsymbol{D}_k$ is simply the identity matrix $\boldsymbol{I}$, while in Newton's method $\boldsymbol{D}_k$ is the inverse of the Hessian $\nabla^2 F(\boldsymbol{x}_k)$. In quasi-Newton methods such as BFGS algorithm ([31]) and its limited memory version ([23]), $\boldsymbol{D}_k$ is an iterative approximation of the inverse Hessian. Since $\boldsymbol{D}_k$ is positive definite, $\boldsymbol{d}_k$ is a descent direction. Moreover, we have the following property:

Property 8 ([2]) Let $\{\boldsymbol{x}_k\}$ a sequence generated by $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$ where $\boldsymbol{d}_k$ is given by (61). If the set $\{\boldsymbol{D}_k, k = 1, \ldots, K\}$ has a positive bounded spectrum, then the direction sequence $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$.

Then, according to Theorem 1, the descent algorithm converges in the sense $\lim_{k\to\infty} \|\boldsymbol{g}_k\| = 0$.

6.2.2 Truncated versions

Let $\boldsymbol{H}_k$ a SPD approximation of the Hessian of F. Thus, a good choice would be to take the preconditioner $\boldsymbol{D}_k = \boldsymbol{H}_k^{-1}$ in (61). However, the calculation of the exact inverse of $\boldsymbol{H}_k$ may be prohibitive, especially when the dimension n is large. One may have to be satisfied with only an approximate solution obtained by using an iterative method. This approach is used in the truncated Newton (TN) algorithm ([29]) where the search direction is computed by applying the conjugate gradient (CG) method to the Newton equations. Here, we consider the more general case when $\boldsymbol{d}_k$ results from CG iterations solving approximately the linear system $\boldsymbol{H}_k \boldsymbol{d} = -\boldsymbol{g}_k$, which will be refered as truncated pseudo-Newton (TPN) algorithms. Then, we have the following property:

Property 9 Let $\{\boldsymbol{x}_k\}$ a sequence generated by $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$ where $\boldsymbol{d}_k$ results from I_k CG iterations on the system $\boldsymbol{H}_k \boldsymbol{d} = -\boldsymbol{g}_k$. If the set $\{\boldsymbol{H}_k, k = 1, \ldots, K\}$ has a positive bounded spectrum, then the direction sequence $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$.

Proof According to [8, Th.A.1] and [8, Lem.A.2], there exist positive constants $\tau, \mathscr{T}$ so that

$$\boldsymbol{g}_k^T \boldsymbol{d}_k \leqslant -\tau \|\boldsymbol{g}_k\|^2 \tag{62}$$

and

$$\|\boldsymbol{d}_k\| \leqslant \mathscr{T} \|\boldsymbol{g}_k\| \tag{63}$$

According to [2, Chap.1], (62) and (63) are sufficient conditions to ensure that $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$. □

Property 9 is extended to the case when the linear system is solved using preconditioned CG (PCG) iterations:

Corollary 3 *Let $\{\boldsymbol{x}_k\}$ a sequence generated by $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$ where $\boldsymbol{d}_k$ results from I_k PCG iterations on the system $\boldsymbol{H}_k \boldsymbol{d} = -\boldsymbol{g}_k$ preconditioned with $\boldsymbol{M}_k$. If $\{\boldsymbol{H}_k, k = 1, \ldots, K\}$ and $\{\boldsymbol{M}_k, k = 1, \ldots, K\}$ have a positive bounded spectrum, then the direction sequence $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$.*

Proof Let $\boldsymbol{C}_k$ such that $\boldsymbol{M}_k = \boldsymbol{C}_k^T \boldsymbol{C}_k$. Solving $\boldsymbol{H}_k \boldsymbol{d} = -\boldsymbol{g}_k$ with PCG preconditioned by $\boldsymbol{M}_k$ amounts to compute vector $\boldsymbol{d}$ such that

$$\boldsymbol{C}_k^{-T} \boldsymbol{H}_k \boldsymbol{C}_k \hat{\boldsymbol{d}} = -\boldsymbol{C}_k^{-T} \boldsymbol{g}_k \tag{64}$$

$$\hat{\boldsymbol{d}} = \boldsymbol{C}_k \boldsymbol{d} \tag{65}$$

using CG iterations ([31]). According to [8, Th.A.1] and [8, Lem.A.2], there exist positive constants $\tau', \mathscr{T}'$ so that

$$(\boldsymbol{C}_k^{-T} \boldsymbol{g}_k)^T \hat{\boldsymbol{d}}_k \leqslant -\tau' \|\boldsymbol{C}_k^{-T} \boldsymbol{g}_k\|^2 \tag{66}$$

and

$$\|\hat{\boldsymbol{d}}_k\| \leqslant \mathscr{T}' \|\boldsymbol{C}_k^{-T} \boldsymbol{g}_k\|. \tag{67}$$

Using (65),

$$(\boldsymbol{C}_k^{-T} \boldsymbol{g}_k)^T \hat{\boldsymbol{d}}_k = \boldsymbol{g}_k^T \boldsymbol{d}_k. \tag{68}$$

Moreover, according to the boundness assumption on the spectrum of $\{\boldsymbol{M}_k, k = 1, \ldots, K\}$,

$$-\|\boldsymbol{C}_k^{-T} \boldsymbol{g}_k\|^2 \leqslant -\frac{1}{\nu_2^{\mathscr{M}}} \|\boldsymbol{g}_k\|^2, \tag{69}$$

$$\|\boldsymbol{C}_k^{-T} \boldsymbol{g}_k\| \leqslant \frac{1}{\sqrt{\nu_1^{\mathscr{M}}}} \|\boldsymbol{g}_k\|, \tag{70}$$

$$\sqrt{\nu_1^{\mathscr{M}}} \|\boldsymbol{d}_k\| \leqslant \|\boldsymbol{C}_k \boldsymbol{d}_k\| = \|\hat{\boldsymbol{d}}_k\|, \tag{71}$$

where $(\nu_1^{\mathscr{M}}, \nu_2^{\mathscr{M}}) > 0$ denote the spectral bounds of $\{\boldsymbol{M}_k\}$. Thus, (62) and (63) hold with $\tau = \tau' \frac{1}{\nu_1^{\mathscr{M}}}$ and $\mathscr{T} = \mathscr{T}' \frac{1}{\nu_2^{\mathscr{M}}}$, hence the result using the gradient related sufficient condition in [2, Chap.1]. $\square$

As a conclusion, the convergence of both TPN-CG and TPN-PCG holds, when the proposed line seach is used, according to Theorem 1.

6.2.3 Feasible directions methods for constrained optimization

Consider the constrained problem:

$$\text{minimize } \; F(\boldsymbol{x}) \; \text{ subject to } \; \boldsymbol{x} \in \mathscr{D}$$

where $\mathscr{D}$ is a nonempty, closed, and convex set. Let us examine the convergence properties of algorithms belonging to the class of feasible direction methods.

Definition 3 ([2])
Given a feasible vector $\boldsymbol{x}$, a feasible direction at $\boldsymbol{x}$ is a vector $\boldsymbol{d} \neq 0$ such that $\boldsymbol{x} + \alpha \boldsymbol{d}$ is feasible for all sufficiently small $\alpha > 0$.

Starting with $\boldsymbol{x}_0 \in \mathscr{D}$, the method generates a sequence of feasible vectors according to

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$$

where $\alpha_k \in (0,1]$ and $\boldsymbol{d}_k$ is a feasible direction that can be written in the form

$$\boldsymbol{d}_k = \boldsymbol{x}'_k - \boldsymbol{x}_k \tag{72}$$

with

$$\boldsymbol{x}'_k \in \mathscr{D}, \quad \boldsymbol{g}_k^T(\boldsymbol{x}'_k - \boldsymbol{x}_k) < 0.$$

Convergence analysis of feasible direction methods is very close to that of descent direction methods in the unconstrained case. In particular, we have the following property:

Property 10 ([2]) Let $\{\boldsymbol{d}_k\}$ generated by (72) with $\boldsymbol{x}'_k$ given either by:

- conditionnal gradient

$$\boldsymbol{x}'_k = \arg\min_{\boldsymbol{x}\in\mathscr{D}} \boldsymbol{g}_k^T(\boldsymbol{x} - \boldsymbol{x}_k) \tag{73}$$

- gradient projection with constant parameter $s > 0$

$$\boldsymbol{x}'_k = \mathscr{P}_{\mathscr{D}}[\boldsymbol{x}_k - s\boldsymbol{g}_k] \tag{74}$$

- scaled gradient projection with constant parameter $s > 0$ and scaling matrices $\{\boldsymbol{D}_k\}$ with bounded spectrum

$$\boldsymbol{x}'_k = \arg\min_{\boldsymbol{x}\in\mathscr{D}} \left\{ \boldsymbol{g}_k^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2s}(\boldsymbol{x} - \boldsymbol{x}_k)^T \boldsymbol{D}_k (\boldsymbol{x} - \boldsymbol{x}_k) \right\} \tag{75}$$

In all these cases, the direction sequence $\{\boldsymbol{d}_k\}$ is gradient related to $\{\boldsymbol{x}_k\}$.

Thus, Theorem 1 implies the convergence of the constrained optimization algorithms defined by (73), (74) and (75), respectively, in conjunction with the proposed line search.

6.3 Convergence of conjugate gradient methods

This section discusses the convergence of the nonlinear conjugate gradient algorithm (NLCG) defined by the following recurrence

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k \\ \boldsymbol{c}_{k+1} &= -\boldsymbol{g}_{k+1} + \beta_{k+1} \boldsymbol{d}_k \\ \boldsymbol{d}_{k+1} &= -\boldsymbol{c}_{k+1} \operatorname{sign}(\boldsymbol{g}_{k+1}^T \boldsymbol{c}_{k+1}) \end{aligned} \tag{76}$$

for some conjugacy formulas.

6.3.1 Methods with $g_k^T y_{k-1}$ in the numerator of β_k

Let us consider the conjugacy formulas of the form ([7]):

$$\beta_0 = 0, \quad \beta_k = \beta_k^{\mu_k,\omega_k} = g_k^T y_{k-1}/D_k, \quad \forall k > 0 \tag{77}$$

with

$$D_k = (1-\mu_k-\omega_k)\|g_{k-1}\|^2 + \mu_k d_{k-1}^T y_{k-1} - \omega_k d_{k-1}^T g_{k-1}$$

$$y_{k-1} = g_k - g_{k-1}$$

$$\mu_k \in [0,1], \quad \omega_k \in [0, 1-\mu_k]$$

Expression (77) covers the following conjugate gradient methods:

$\beta_k^{1,0} = g_k^T y_{k-1}/d_{k-1}^T y_{k-1}$	Hestenes-Stiefel (HS)
$\beta_k^{0,0} = g_k^T y_{k-1}/\|g_{k-1}\|^2$	Polak-Ribière-Polyak (PRP)
$\beta_k^{0,1} = -g_k^T y_{k-1}/d_{k-1}^T g_{k-1}$	Liu-Storey (LS)

The following convergence result holds:

Theorem 2 *Let Assumption 1 and 5 hold. The NLCG algorithm is convergent in the sense* $\liminf_{k\to\infty} g_k = 0$ *when* α_k *is defined by* (16) *and* β_k *is chosen according to the PRP and LS methods, and more generally for* $\mu_k = 0$ *and* $\omega_k \in [0,1]$. *Moreover, if Assumption 3 holds, then we have* $\liminf_{k\to\infty} g_k = 0$ *in all cases.*

Proof We have previously established:

- the inequality (33) on α_k^1
- the stepsize minorization (44) $\alpha_k \leqslant c_J^{\max}\alpha_k^1$
- the stepsize majorization (52) $0 \leqslant c_{\min}\alpha_k^1 \leqslant \alpha_k$
- the fulfillment of Zoutendijk condition (58)

Thus, the proof of Theorem 2 is identical to that developped in [22, Part 4]. This result can be viewed as an extension of [22, Th. 4.1] for a new form of tangent majorant.

□

6.3.2 Other conjugacy formulas

Let consider the following conjugacy formulas:

$\beta_k = \max(g_{k+1}^T(g_{k+1}-g_k)/\|g_k\|, 0)$	modified Polak-Ribière-Polyak (PRP+)
$\beta_k = \|g_{k+1}\|^2/\|g_k\|^2$	Fletcher-Reeves (FR)
$\beta_k = \|g_{k+1}\|^2/d_k^T(g_{k+1}-g_k)$	Dai-Yuan (DY)

The convergence of the CG algorithm with these conjugacy formulas is obtained under an additional assumption on the tangent majorant.

Theorem 3 *Let α_k be defined by the recurrence* (16)*. According to Assumptions 1 and 5, if for all $j \in \{0,\dots,J-1\}$,* (30) *holds, then we have convergence in the sense* $\liminf_{k\to\infty} \boldsymbol{g}_k = 0$ *for the PRP+ and FR methods. Moreover, under Assumption 2, we have convergence in the same sense for the DY method.*

Proof We will prove by recurrence on k that $\boldsymbol{d}_k$ is a sufficient descent direction for F, i.e., there exists $\eta > 0$ such that

$$\boldsymbol{g}_k^T \boldsymbol{d}_k \leqslant -\eta \|\boldsymbol{g}_k\|^2. \tag{78}$$

Let $\boldsymbol{x}_k \in \mathscr{V}$ and let $\boldsymbol{d}_k$ a sufficient descent direction. Let us prove that $\boldsymbol{d}^{k+1}$ is a sufficient descent direction. According to Lemma 5, (30) implies that $\dot{f}(\alpha^j) < 0$ for all j. Thus $\boldsymbol{g}_{k+1}^T \boldsymbol{d}_k \leqslant 0$. From (76),

$$\boldsymbol{g}_{k+1}^T \boldsymbol{c}_{k+1} = -\|\boldsymbol{g}_{k+1}\|^2 + \beta_{k+1} \boldsymbol{g}_{k+1}^T \boldsymbol{d}_k$$

Let us consider the case of FR and PRP+ methods:

$$\beta_k^{\mathrm{FR}} = \frac{\|\boldsymbol{g}_{k+1}\|^2}{\|\boldsymbol{g}_k\|^2} \geqslant 0 \tag{79}$$

$$\beta_k^{\mathrm{PRP+}} = \max(\beta_k^{\mathrm{PRP}}, 0) \geqslant 0 \tag{80}$$

Thus, $\boldsymbol{g}_{k+1}\boldsymbol{c}_{k+1} \leqslant -\|\boldsymbol{g}_{k+1}\|^2$, so $\boldsymbol{d}^{k+1} = \boldsymbol{c}^{k+1}$ is a sufficient descent direction. Now, consider the case of DY conjugacy:

$$\beta_k^{\mathrm{DY}} = \frac{\|\boldsymbol{g}_{k+1}\|^2}{\boldsymbol{d}_k^T (\boldsymbol{g}_{k+1} - \boldsymbol{g}_k)}$$

The conjugacy parameter takes the sign of $\boldsymbol{d}_k^T(\boldsymbol{g}_{k+1} - \boldsymbol{g}_k)$. Under Assumption 2 and given (76), the convexity of F leads to

$$|\boldsymbol{g}_{k+1}^T \boldsymbol{d}_k| \leqslant |\boldsymbol{g}_k^T \boldsymbol{d}_k| \tag{81}$$

Since $\boldsymbol{d}_k$ is a descent direction, $\beta_k^{\mathrm{DY}} \geqslant 0$, so $\boldsymbol{d}^{k+1} = \boldsymbol{c}^{k+1}$ is a sufficient descent direction. Then, (78) holds for all k for FR, DY and PRP+ methods. Finally, according to [16, Th. 4.2, Th. 5.1], Property 7 and (78) yield the convergence of the PRP+, FR and DY methods. □

7 Experimental results

This section presents three application examples illustrating the practical efficiency of the proposed line search procedure. The examples are chosen from the field of image and signal processing.

7.1 Image reconstruction under Poisson noise

We consider a simulated positron emission tomography (PET) ([32]) reconstruction problem. The measurements in PET are modeled as Poisson random variables:

$$\boldsymbol{y} \sim Poisson(\boldsymbol{H}\boldsymbol{x}+\boldsymbol{r})$$

where the ith entry of $\boldsymbol{x}$ represents the radioisotope amount in pixel n and $\boldsymbol{H}$ is the projection matrix whose elements H_{mn} model the contribution of the nth pixel to the mth datapoint. The components of $\boldsymbol{y}$ are the counts measured by the detector pairs and $\boldsymbol{r}$ models the background events (scattered events and accidental coincidences). The aim is to reconstruct the image $\boldsymbol{x} \geqslant 0$ from the noisy measurements $\boldsymbol{y}$.

7.1.1 Objective function

According to the noise statistics, the neg-log-likelihood of the emission data is

$$J(\boldsymbol{x}) = \sum_{m=1}^{M} \left([\boldsymbol{H}\boldsymbol{x}]_m + r_m - y_m \log([\boldsymbol{H}\boldsymbol{x}]_m + r_m)\right).$$

The penalization term resulting from modelling the pixel intensity distribution using a gamma-mixture density is ([17]):

$$R(\boldsymbol{x}) = -\sum_{n=1}^{N} \left((a_n - 1)\log x_n - \frac{a_n}{b_n} x_n\right).$$

Here, the parameters $a_n > 1$ and $b_n \geqslant 0$ of the gamma priors are assumed to take known values[1]. The estimated image is the minimizer of the following objective function

$$F(\boldsymbol{x}) = J(\boldsymbol{x}) + R(\boldsymbol{x}). \tag{82}$$

The first part of the criterion implies the presence of a logarithmic barrier in J. The second part corresponds to a gamma-mixture prior that enforces positivity into account and favors the clustering of pixel intensities. It induces a second type of log barrier, at the boundary of the positive orthant. A classical approach for solving the optimization problem is to use the NLCG algorithm ([17]) with the Moré and Thuente's (MT) line search procedure ([26]). We propose to compare the performance of the algorithm when our MM line search procedure is used.

7.1.2 Optimization strategy

The NLCG algorithm is employed with PRP+ conjugacy. The convergence of the algorithm with the proposed line search is established in Theorem 3 under Assumptions 1, 5 and condition (30). Let $J = P + B$ with

$$B(\boldsymbol{x}) = \sum_{m=1}^{M} -y_m \log([\boldsymbol{H}\boldsymbol{x}]_m + r_m) + \sum_{n=1}^{N} (a_n - 1)\log x_n,$$

[1] Hyperparameters estimation is discussed in ([17]). However, the resulting algorithm does not fall within the application of our convergence theory and the adaptation would require a specific analysis.

and

$$P(\boldsymbol{x}) = \sum_{m=1}^{M} [\boldsymbol{H}\boldsymbol{x}]_m + r_m + \sum_{n=1}^{N} \frac{a_n}{b_n} x_n.$$

It is straightforward that Assumption 1 holds for all $\boldsymbol{x}_0 > 0$. Moreover, Assumption 5 holds for $M(\boldsymbol{x}) = 0$, $\boldsymbol{A} = [Id\ \boldsymbol{H}]^T$ and $\rho = [\mathbf{0}\ \boldsymbol{r}]^T$. Finally, since P is linear, condition (30) reads:

$$0 \leqslant m_p^j(\alpha - \alpha^j), \quad \forall \alpha \geqslant \alpha^j$$

and holds for $m_p^j = M(\boldsymbol{x} + \alpha^j \boldsymbol{d}) = 0$. Theorem 3 does not cover the preconditioned case. However, we have noticed that, in practice, the use of a diagonal preconditioner substantially speeds up the algorithm convergence.

The algorithm is initialized with a uniform positive object and the convergence is checked using the following stopping rule ([31])

$$\|\boldsymbol{g}_k\|_\infty < \varepsilon(1 + |F(\boldsymbol{x}_k)|), \tag{83}$$

where ε is set to 10^{-7}.

7.1.3 Results and discussion

We present a simulated example using data generated with J.A. Fessler's code available at `http://www.eecs.umich.edu/~fessler`. For this simulation, we consider an image $\boldsymbol{x}^o$ of size $N = 128 \times 128$ pixels and $M = 24924$ pairs of detectors. Table 1 summarizes the performance results in terms of iteration number K and computation time T on an Intel Pentium 4, 3.2 GHz, 3 GB RAM. The design parameters are the Wolfe condition constants (c_1, c_2) for the MT method and the number of subiterations J for the MM procedure.

	c_1	c_2	K	T(s)
NLCG-MT	10^{-3}	0.5	97	361
	10^{-3}	0.9	107	337
	10^{-3}	0.99	102	317
	10^{-3}	0.999	102	<u>313</u>
	J		K	T(s)
NLCG-MM	1		96	<u>266</u>
	2		111	464
	5		138	1526
	10		138	3232

Table 1 Comparison between MM and MT line search strategies for a PET reconstruction problem solved with NLCG algorithm, in terms of iteration number K and time T before convergence. Convergence is considered in the sense of (83).

It can be noted that the NLCG algorithm with MM line search (NLCG-MM) requires less iterations than the MT method (NLCG-MT), even when the parameters (c_1, c_2) are optimally chosen. Moreover, NLCG-MM is faster because of a smaller computational cost per iteration. Furthermore, the proposed MM procedure admits a unique tuning parameter, namely the subiteration number J, and the simplest choice $J = 1$ appears the best one.

7.2 Nuclear magnetic resonance reconstruction

We consider a mono-dimensional nuclear magnetic resonance (NMR) reconstruction problem. The NMR decay $s(t)$ associated with a continuous distribution of relaxation constants $x(T)$ is described in terms of a Fredholm integral of the first kind:

$$s(t) = \int_{T_{\min}}^{T_{\max}} x(T)\, k(t,T)\, dT. \tag{84}$$

with $k(t,T) = e^{-t/T}$. In practice, the measured signal $\boldsymbol{s}$ is a set of discrete experimental noisy data points $s_m = s(t_m)$ modeled as

$$\boldsymbol{s} = \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\varepsilon} \tag{85}$$

where $\boldsymbol{K}$ and $\boldsymbol{x}$ are discretized versions of $k(t,T)$ and $x(T)$ with dimensions $M \times N$ and $N \times 1$, and $\boldsymbol{\varepsilon}$ is an additive noise assumed centered white Gaussian. Given $\boldsymbol{s}$, the aim is to determine $\boldsymbol{x} \geqslant 0$. This problem is equivalent to a numerical inversion of the Fredholm integral (84) and is known as very ill-conditioned ([4]).

7.2.1 Objective function

In order to get a stabilized solution, an often used method minimizes the expression

$$F(\boldsymbol{x}) = J(\boldsymbol{x}) + \lambda R(\boldsymbol{x}) \tag{86}$$

under positivity constraints, where J is a fidelity to data term:

$$J(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{s} - \boldsymbol{K}\boldsymbol{x}\|_2^2,$$

and R is an entropic regularization term, e.g., the Shannon entropy measure ([24]):

$$R(\boldsymbol{x}) = \sum_n x_n \ln x_n$$

Moreover, the positivity constraint is implicitely handled because of the barrier property of the entropy function.

7.2.2 Optimization strategy

The TN algorithm is employed for solving (86). The direction $\boldsymbol{d}_k$ is computed by approximately solving the Newton system $\nabla^2 F(\boldsymbol{x}_k)\boldsymbol{d} = -\boldsymbol{g}_k$ using PCG iterations. We propose a preconditioning matrix $\boldsymbol{M}_k$ built as an approximation of the inverse Hessian of F at $\boldsymbol{x}_k$:

$$\boldsymbol{M}_k = \left[\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T + \lambda\, \mathrm{diag}(\boldsymbol{x}_k)^{-1}\right]^{-1},$$

where $\boldsymbol{U}^T\boldsymbol{\Sigma}\boldsymbol{V}$ is a truncated singular value decomposition of $\boldsymbol{K}$ and $\boldsymbol{D} = \boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$. The convergence of the TN algorithm with the proposed line search is established in Theorem 1 using Corollary 3 under Assumptions 1 and 5. The verification of the latter is straightforward for $M(\boldsymbol{x}) = \boldsymbol{K}^T\boldsymbol{K}$, $\boldsymbol{A} = Id$ and $\rho = \mathbf{0}$. The fulfillment of Assumption 1 is more difficult to check since the level set $\mathscr{L}_0$ may contain an element $\boldsymbol{x}$ with zero components, contradicting the gradient Lipshitz assumption. In practice, we initialized the algorithm with $\boldsymbol{x}_0 > 0$ and we never noticed convergence issues in our practical tests. The extension of the convergence results under a weakened version of Assumption 1 remains an open issue in our convergence analysis.

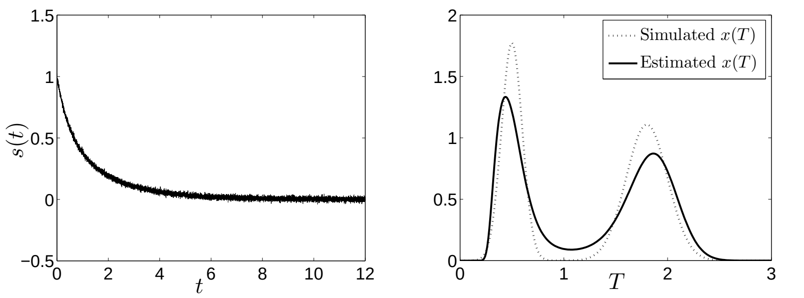
The algorithm is initialized with a uniform positive object and the convergence is checked using (83) with $\varepsilon = 10^{-9}$. Following [29], the PCG iterations are stopped when:

$$\|\nabla F(\boldsymbol{x}_k) + \nabla^2 F(\boldsymbol{x}_k)\boldsymbol{d}_k\| \leqslant 10^{-5}\|F(\boldsymbol{x}_k)\|.$$

We propose to compare the performances of the MM line search and of the interpolation-based MT method [26].

7.2.3 Results and discussion

Let $\boldsymbol{x}(T)$ a distribution to estimate. We consider the resolution of (85) when data $\boldsymbol{s}$ are simulated from $\boldsymbol{x}(T)$ via the NMR model (85) over sampled times $t_m, m = 1,...,10000$, with a SNR of 25 dB (Figure 1). The regularization parameter λ is set to $\lambda = 7,2 \cdot 10^{-4}$ to get the best result in terms of similarity between the simulated and the estimated spectra (in the sense of quadratic error).



(a) Simulated NMR measurement with SNR $= 25dB$

(b) NMR reconstruction with similarity error 8.5%.

Fig. 1 Simulated NMR reconstruction with maximum entropy method

According to Table 2, the TN algorithm with the MM line search performs better than with TN with the best settings for c_1 and c_2. Concerning the choice of the sub-iteration number, it appears that $J = 1$ leads again to the best results in terms of computation time.

	c_1	c_2	K	T(s)
TN-MT	10^{-3}	0.5	34	<u>12</u>
	10^{-3}	0.9	42	13
	10^{-3}	0.99	71	20
	10^{-2}	0.99	71	19
	10^{-2}	0.5	34	13
	10^{-1}	0.99	71	19
	10^{-1}	0.5	34	14
	J		K	T(s)
TN-MM	1		36	<u>8</u>
	2		40	9
	5		40	10
	10		40	14

Table 2 Comparison between MM and MT line search strategies for a maximum entropy NMR reconstruction problem solved with TN algorithm, in terms of iteration number K and time T before convergence. Convergence is considered in the sense of (83).

7.3 Constrained quadratic programming

Let consider the following quadratically constrained quadratic optimization problem

$$\min_{\boldsymbol{x}} \left\{ F_0(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}_0 \boldsymbol{x} + a_0^T \boldsymbol{x} + \rho_0 \right\} \tag{87}$$
$$\text{subject to: } C_i(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}_i \boldsymbol{x} + a_i^T \boldsymbol{x} + \rho_i \geqslant 0,\, i = 1, \ldots, m$$

where $\boldsymbol{A}_i, i = 0, \ldots, m$ are SPD matrices of $\mathbb{R}^{n\times n}$. We propose to solve (87) with the primal interior point algorithm of [3]: for a decreasing sequence of barrier parameters μ, the augmented criterion

$$F_\mu(\boldsymbol{x}) = F_0(\boldsymbol{x}) - \mu \sum_{i=1}^{m} \log C_i(\boldsymbol{x}).$$

is minimized using Newton iterations

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \quad \text{with} \quad \boldsymbol{d}_k = -\nabla^2 F_\mu^{-1}(\boldsymbol{x}_k) \nabla F_\mu(\boldsymbol{x}_k)$$

that are stopped when $(\boldsymbol{d}_k^T \boldsymbol{g}_k)^2 \leqslant 2\varepsilon$.

The stepsize α_k must belong to an interval (α_-, α_+) that corresponds to the definition domain of $F_\mu(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$. Since the constraints are quadratic in $\boldsymbol{x}$, they are also quadratic in α:

$$C_i(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) = Q_i^1 \alpha^2 + Q_i^2 \alpha + Q_i^3$$

with $Q_i^1 = -\frac{1}{2}\boldsymbol{d}_k^T \boldsymbol{A}_i \boldsymbol{d}_k$, $Q_i^2 = -\boldsymbol{x}_k^T \boldsymbol{A}_i \boldsymbol{d}_k + a_i^T \boldsymbol{d}_k$ and $Q_i^3 = -\frac{1}{2}\boldsymbol{x}_k^T \boldsymbol{A}_i \boldsymbol{x}_k + a_i^T \boldsymbol{x}_k + \rho_i$. As a consequence, α_- and α_+ can be computed exactly for any $(\boldsymbol{x}_k, \boldsymbol{d}_k)$. For example, α_+ is the smallest positive root of the concave polynomes $C_i(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$. In [3], the stepsize strategy is based on backtracking. Starting with the feasible step $\alpha = 0.99\,\alpha_+$, the stepsize is reduced until it fulfills the first Wolfe condition (12). As an alternative in the context of interior point methods, a *damped Newton* approach is developped in [30] to minimize the augmented criterion F_μ. The Newton direction $\boldsymbol{d}_k$ is damped by a factor $\alpha_k \in (0,1]$ ensuring that $\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$ is feasible and that the criterion decreases by a minimal fixed amount. The damping factor is given by

$$\alpha_k = \frac{1}{1 + \|\boldsymbol{d}_k\|_{\boldsymbol{x}_k}}$$

where $\|\cdot\|_{\boldsymbol{x}}$ is the Hessian norm defined by $\|\boldsymbol{u}\|_{\boldsymbol{x}} = \sqrt{\boldsymbol{u}^T \nabla^2 F_\mu(\boldsymbol{x}) \boldsymbol{u}}$.

The convergence properties of this interior point algorithm are based on the self concordancy of F_μ ([30]). Our aim here is only to evaluate the practical relevance of the MM line search when it is used instead of the backtracking and the damping procedures.

7.4 Results and discussion

In order to analyse the performance of the interior point algorithm, we apply it onto 50 problems with $\boldsymbol{A}_i, \rho_i$ and a_i generated randomly taking $n = 400$, $m = 200$ as in [20]. $\boldsymbol{x}$ is initialized in the constrained domain $\mathscr{C}$. The barrier parameter μ is initially set to 1 and decreases following a geometric series of ratio 0.2. The algorithm is stopped when $\mu \leqslant \mu_{\min}$. Table 3 reports the performances of the interior point algorithm for the different line search procedures using $c_1 = 0.01$ and $J = 1$.

	Backtracking	Damping	MM
K	273 ± 27	135 ± 4	64 ± 3
$T(s)$	5637 ± 1421	465 ± 26	225 ± 8

Table 3 Comparison between different line search strategies for the interior point algorithm over 50 random quadratic programming problems. K denotes the sum of inner iterations and T the time before convergence, with tolerance parameters $\mu_{\min} = 10^{-8}$ and $\varepsilon = 10^{-5}$. The results are given in terms of mean and standard deviation.

It can be noted that the interior point algorithm with MM line search requires less iterations than the backtracking and damped Newton approaches. Moreover, even if the MM procedure requires the exact computation of (α_-, α_+), it is faster than the two other approaches. It can also be remarked that the damping strategy is dedicated to the particular case when $\boldsymbol{d}$ is the Newton direction. Therefore, it must be modified when the minimization of F_μ is obtained by means of other algorithms (see [20] for the conjugate gradient case). On the contrary, the proposed line search can be applied independently of the descent algorithm used. To conclude, the MM procedure seems an efficient alternative to line search strategies widely used in primal interior point algorithms.

8 Conclusion

This paper extends the line search strategy of [22] to the case of criteria containing barrier functions, by proposing a non-quadratic majorant approximation of the function in the line search direction. This majorant has the same form as the one proposed in [27], whereas the latter follows an interpolation-based approach. However, in the majorization-based approach, the construction of the approximation is easier and its minimization leads to an analytical stepsize formula, guaranteeing the convergence of several descent algorithms. Moreover, numerical experiments indicate that this approach outperforms standard line search methods based on backtracking, damping or cubic interpolation.

Two extensions of this work are envisaged. On the one hand, the case of nonlinear constraints can be handled by using the procedure described in [27]. On the other hand, the analysis can be performed for additionnal forms of barrier functions such as cross-entropy ([33]) or inverse function ([9]).

References

1. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods, 2nd edn. Athena Scientific, Belmont, MA (1996)
2. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Belmont, MA (1999)
3. Boyd, S., Vandenberghe, L.: Convex Optimization, 1st edn. Cambridge University Press, New York (2004)
4. Butler, J.P., Reeds, J.A., Dawson, S.V.: Estimating solutions of first kind integral equations with non-negative constraints and optimal smoothing. SIAM Journal on Numerical Analysis **18**(3), 381–397 (1981)
5. Charbonnier, P., Blanc-Féraud, L., Aubert, G., Barlaud, M.: Deterministic edge-preserving regularization in computed imaging. IEEE Transactions on Image Processing **6**, 298–311 (1997)
6. Conn, R.A., Gould, N.I.M., Toint, P.L.: Trust-Region Methods. MPS-SIAM Series on Optimization. Society for Industrial Mathematics (1987)
7. Dai, Y., Yuan, Y.: A three-parameter family of nonlinear conjugate gradient methods. Mathematics of Computation **70**, 1155–1167 (2001)
8. Dembo, R.S., Steihaug, T.: Truncated-newton methods algorithms for large scale unconstrained optimization. Mathematical Programming **26**, 190–212 (1983)
9. Den Hertog, D., Roos, C., Terlaky, T.: Inverse barrier methods for linear programming. Revue française d'automatique, informatique, recherche opérationnelle **28**(2), 135–163 (1994)
10. Dixon, L.C.W.: Conjugate directions without linear searches. IMA Journal of Applied Mathematics **11**(3), 317–328 (1973)
11. Doyle, M.: A barrier algorithm for large nonlinear optimization problems. PhD thesis, University of Stanford (2003). `www.stanford.edu/group/SOL/dissertations/maureenthesis.pdf`
12. Fessler, J.A., Booth, S.D.: Conjugate-gradient preconditiong methods for shift-variant PET image reconstruction. IEEE Transactions on Image Processing **8**(5), 688–699 (1999)
13. Figueiredo, M., Bioucas-Dias, J., Nowak, R.: Majorization-minimization algorithms for wavelet-based image restoration. IEEE Transactions on Image Processing **16**(12), 2980–2991 (2007)
14. Fletcher, R.: On the Barzilai-Borwein method. In: L. Qi, K. Teo, X. Yang (eds.) Optimization and Control with Applications, vol. 96, pp. 235–256. Springer (2005)
15. Forsgren, A., Gill, P., Wright, M.: Interior methods for nonlinear optimization. SIAM Review **44**(4), 525–597 (2002)
16. Hager, W.W., Zhang, H.: A survey of nonlinear conjugate gradient methods. Pacific Journal of Optimization **2**(1), 35–58 (2006)
17. Hsiao, I.T., Rangarajan, A., Gindi, G.: Joint-MAP Bayesian tomographic reconstruction with a gamma-mixture prior. IEEE Transactions on Image Processing **11**(12), 1466–1477 (2002). DOI 10.1109/TIP.2002.806254
18. Hunter, D.R., L., K.: A tutorial on MM algorithms. The American Statistician **58**(1), 30–37 (2004)
19. Jacobson, M., Fessler, J.: An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms. IEEE Transactions on Image Processing **16**(10), 2411–2422 (2007)
20. Ji, H., Huang, M., Moore, J., Manton, J.: A globally convergent conjugate gradient method for minimizing self-concordant functions with application to constrained optimisation problems. In: American Control Conference, pp. 540–545 (2007). DOI 10.1109/ACC.2007.4282797

21. Labat, C., Idier, J.: Convergence of truncated half-quadratic and Newton algorithms, with application to image restoration. Technical report, IRCCyN (2007)
22. Labat, C., Idier, J.: Convergence of conjugate gradient methods with a closed-form stepsize formula. Journal of Optimization Theory and Applications **136**(1), 43–60 (2008)
23. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical Programming **45**(3), 503–528 (1989)
24. Mariette, F., Guillement, J.P., Tellier, C., Marchal, P.: Continuous relaxation time distribution decomposition by MEM. Signal Treatment and Signal Analysis in NMR pp. 218–234 (1996)
25. Merabet, N.: Global convergence of a memory gradient method with closed-form step size formula. Discrete and Continuous Dynamical Systems pp. 721–730 (2007)
26. Moré, J.J., Thuente, D.J.: Line search algorithms with guaranteed sufficient decrease. ACM Transactions on Mathematical Software **20**(3), 286–307 (1994)
27. Murray, W., Wright, M.H.: Line search procedures for the logarithmic barrier function. SIAM Journal on Optimization **4**(2), 229–246 (1994)
28. Narkiss, G., Zibulevsky, M.: Sequential subspace optimization method for large-scale unconstrained problems. Technical Report 559, Israel Institute of Technology (2005)
29. Nash, S.G.: A survey of truncated-Newton methods. Journal of Computational and Applied Mathematics **124**, 45–59 (2000)
30. Nesterov, Y., Nemirovskii, A.: Interior point polynomial algorithms in convex programming. No. 13 in Studies in Applied and Numerical Mathematics. SIAM, Philadelphia, Pennsylvania (1994)
31. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer-Verlag, New York, NY (1999)
32. Ollinger, J.M., Fessler, J.A.: Positron-emission tomography. IEEE Signal Processing Magazine **14**(1), 43–55 (1997)
33. O'Sullivan, J.: Roughness penalties on finite domains. IEEE Transactions on Image Processing **4**(9) (1995)
34. Roos, C., Terlaky, T., Vial, J.: Interior Point Methods for Linear Optimization, 2nd edn. Springer-Verlag, New York, NY (2006)
35. Shi, Z.J.: Convergence of line search methods for unconstrained optimization. Applied Mathematics and Computations **157**, 393–405 (2004)
36. Skilling, J., Bryan, R.K.: Maximum entropy image reconstruction: General algorithm. Monthly Notices of the Royal Astronomical Society **211**, 111–124 (1984)
37. Sun, J., Zhang, J.: Global convergence of conjugate gradient methods without line search. Annals of Operations Research **103**, 161–173 (2001)